A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin

Wagner, Andreas

Abstract: Networks of evolving genotypes can be constructed from the worldwide time-resolved genotyping of pathogens like influenza viruses. Such genotype networks are graphs where neighbouring vertices (viral strains) differ in a single nucleotide or amino acid. A rich trove of network analysis methods can help understand the evolutionary dynamics reflected in the structure of these networks. Here, I analyse a genotype network comprising hundreds of influenza A (H3N2) haemagglutinin genes. The network is rife with cycles that reflect non-random parallel or convergent (homoplastic) evolution. These cycles also show patterns of sequence change characteristic for strong and local evolutionary constraints, positive selection and mutation-limited evolution. Such cycles would not be visible on a phylogenetic tree, illustrating that genotype network analysis can complement phylogenetic analyses. The network also shows a distinct modular or community structure that reflects temporal more than spatial proximity of viral strains, where lowly connected bridge strains connect different modules. These and other organizational patterns illustrate that genotype networks can help us study evolution in action at an unprecedented level of resolution.

DOI: https://doi.org/10.1098/rspb.2013.2763

A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) hemagglutinin

Andreas Wagner [1,2,3]

[1] *Institute of Evolutionary Biology and Environmental Sciences, Bldg. Y27, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland*

[2] *The Swiss Institute of Bioinformatics, Bioinformatics, Quartier Sorge, Batiment Genopode, 1015 Lausanne, Switzerland.*

[3] *The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

Keywords: evolution, networks, influenza, cycles, homoplasy

Abstract

Networks of evolving genotypes can be constructed from the world-wide time-resolved genotyping of pathogens like influenza viruses. Such genotype networks are graphs where neighboring vertices (viral strains) differ in a single nucleotide or amino acid. A rich trove of network analysis methods can help understand the evolutionary dynamics reflected in the structure of these networks. Here I analyze a genotype network comprising hundreds of influenza A (H3N2) hemagglutinin genes. The network is rife with cycles that reflect non-random parallel or convergent (homoplastic) evolution. These cycles also show patterns of sequence change characteristic for strong and local evolutionary constraints, positive selection, and mutation-limited evolution. Such cycles would not be visible on a phylogenetic tree, illustrating that genotype network analysis can complement phylogenetic analyses. The network also shows a distinct modular or community structure that reflects temporal more than spatial proximity of viral strains, where lowly connected bridge strains connect different modules. These and other organizational patterns illustrate that genotype networks can help us study evolution in action at an unprecedented level of resolution.

Introduction

The human influenza A virus causes up to half a million deaths annually and infects more than 5 percent of the world's population (Hayward et al. 2014; WHO 2009). The virus evades the human immune system through antigenic change, requiring costly regular updates of influenza vaccines (Carrat & Flahault 2007; Smith et al. 2004). A major target of the immune response is the viral hemagglutinin (HA) protein, a surface glycoprotein that enables the virus to bind and enter host cells via sialic acid residues on cell surface receptors (Wagner et al. 2002). HA is a homotrimeric membrane protein whose monomers form three globular heads, as well as a helical coil that resides in the membrane (Wilson et al. 1981). The globular heads contain epitopes – those parts of a foreign molecule recognized by the immune system – which can be bound by antibodies that prevent HA's interaction with host cells (Wiley et al. 1981). Influenza A subtypes are classified according to the their variants of hemagglutinin and neuraminidase, a viral protein that is important to release viruses from the cell surface. The H3N2 subtype (hemagglutinin subtype 3, neuraminidase subtype 2) dominates the seasonal flu that recurs annually in temperate regions. Intense monitoring of influenza epidemiology and evolution (Hayward et al. 2014) make this virus ideal for novel, data-intensive approaches to understand the evolutionary dynamics of pathogens, such as the framework of genotype networks.

Genotype networks are graphs whose nodes are genotypes with the same broadly defined phenotype. This phenotype could be as coarse-grained as "being viable", or as fine-grained as the enzymatic activity or catalytic site conformation of a protein. Two genotypes are neighbors and connected by an edge in such a network if they differ minimally, e.g., in a single nucleotide or amino acid. Genotype networks and their structure can shed new light on many long-standing problems in evolutionary biology, such as how new evolutionary adaptations originate (Wagner 2011). Thus far, genotype networks have mainly been characterized in systems where phenotypes need to be predicted computationally from genotypes (Ciliberti et al. 2007; Lipman & Wilbur 1991; Rodrigues & Wagner 2009; Schuster et al. 1994). High-throughput genotyping technologies can alleviate this limitation and help build genotype networks from experimental data, by characterizing many closely related genotypes and their phenotypes (Bershtein et al. 2006; Hayden et al. 2011; Hietpas et al. 2011; Pybus & Rambaut 2009; Romero & Arnold 2009). Casting such data in the form of a network immediately makes many analytical tools from graph theory available (Harary 1969). Their use has contributed to fields as different as ecology, systems biology, and the social sciences (Bascompte & Jordano 2007; Bascompte et al. 2006; Cohen & Briand 1984; Maslov & Sneppen 2002; Milo et al. 2002; Newman 2006; Onnela et al. 2007). For example, they can help identify "modules" of cooperating molecules, interactions in ecological networks that affect their stability, or network properties that can influence the spreading of traits, such as innovations or diseases (Bascompte & Jordano 2007; Colizza et al. 2006; Newman 2006).

The conventional way to study evolutionary processes with genotypic data (but see (Bull et al. 2008; Rozenfeld et al. 2008)) is to construct phylogenetic trees that reflect the evolutionary relationships among genes, individuals, or species (Felsenstein 2004). Different genotypes are leaf nodes on such a tree. Internal (non-leaf) nodes correspond to usually extinct ancestors. Modern phylogenetic methods permit a probabilistic reconstruction of such ancestors, in the sense that they can compute a probability that a given genotype was the true common ancestor of the actually observed leaf nodes. They allow

one to choose the most likely such genotype as the best candidate common ancestor (Felsenstein 2004; Guindon & Gascuel 2003). In both phylogenetic trees and genotype networks, the distance between two sequences reflects their evolutionary relatedness, but genotype networks differ from phylogenetic trees in at least two respects. First, when constructed from a population sample, they do not require reconstruction of ancestors, but contain these ancestors as their internal nodes. Second, as opposed to trees – which are by definition acyclic graphs – they can contain cycles, paths of edges that return to their origin, which can indicate unusual patterns of evolution.

I here construct a genotype network from a large data set of viable influenza genotypes to illustrate some of its applications to evolutionary genetics. Specifically, I construct a genotype network based on more than 1000 influenza A hemagglutinin sequences isolated from H3N2 viruses between 2002 and 2007 (Russell et al. 2008) to illustrate how graph theoretical concepts and methods can shed light on the evolution of this pathogen.

Results

Structure of the protein genotype network. The protein-based genotype network is a graph whose nodes are hemagglutinin protein sequences from different viral strains. Edges connect two sequences if they differ in a single amino acid. The network based on the data set used here (Russell et al. 2008) has 1565 sequences, 33763 edges, and is organized into 284 components – subgraphs where any two sequences can be connected via a path of edges (Figure S1). After removal of all but one identical sequences from this network, the non-redundant network still contains 742 sequences, with 539 edges between them, which are partitioned among 265 connected components of various sizes (Figure 1a). Most components (238) consist of a single isolated sequence, 27 components contain two or more sequences, and the single largest "giant" component (Bollobás 1985) comprises 246 sequences connected by 285 edges. Figure 1b shows this largest component. The distribution of sequence degrees – a sequence's number of neighbors – is highly skewed towards sequences with few neighbors (Figure 1c). The network is disassortative (Newman 2002), meaning that sequences with high degree tend to be neighbors of sequences with low degree (Figure S2). It has a highly modular or community structure (Figure S3), where a strain's membership in a module is most significantly associated with its year of isolation (Supplementary Results, Figure S4). An analogous network, based on DNA sequences rather than protein sequences, is highly fragmented and thus less informative (Supplementary Results, Figure S5).

*Abundant cycles in the genotype network cannot be explained by random homoplasy***.** Whereas cycles cannot occur in a phylogenetic tree by definition (Felsenstein 2004), they can occur in genotype networks. Such cycles can reflect constrained evolution and in particular homoplasy, parallel or convergent evolution, where two sequences do not diverge or even become more similar over time. The HA protein genotype network contains remarkably many cycles. Specifically, among the 504 sequences in the network that are not isolated, 122 (24 percent) form part of a cycle. All of those cycles are contained in the six largest components, in which 28.6 percent of sequences form part of a cycle. In the largest component itself, 79 of 246 sequences (32.1 percent) are contained in cycles. To better visualize the extent of cycles, one can display the so-called 2-core of the genotype network (Figure 2a), defined as the largest subgraph where every sequence has at least two neighbors. All 122 sequences of this 2-core

form part of a cycle. The 2-core thus reveals the extended cyclic structure of the protein genotype network.

Detailed examination shows that all cycles in the network are decomposable into triangles and squares (Tables S1 and S2). First, the network contains 24 triangles that involve only 48 sequences, implying that many triangles share sequences and edges (See also Figure 2a). Second, the network contains 40 squares that involve a total of 94 sequences which are shared among squares. Twenty sequences are shared between triangles and squares, implying that 28 sequences in a cycle occur only in a triangle and 74 sequences in a cycle occur only in a square (Figure 2b). The network contains many longer cycles of five or more edges, but none of them is an elementary cycle (Figure S6) – all of them can be decomposed into triangles and squares.

I next developed an algorithm (see Methods) to ask whether the cycles in the genotype network could have arisen by chance alone, i.e., from the limited amount of homoplasy to be expected in independently evolving sequences. With this algorithm, I created 1000 random genotype networks, and counted the number of sequences involved in cycles in them. Among 1000 random genotype networks of the same size as the largest component of the HA genotype network, 999 contained no cycle at all, one contained a triangle, and none contained a square. Not a single network contained more than a triangle or square. Similarly, 1000 random networks of the same size as the second- and third-largest component did not contain a single cycle. The number of cycles observed in the HA genotype network cannot be explained by chance alone.

Global evolutionary constraints cannot explain abundant cycles. Because the HA gene is important to the viral life cycle, its evolution is constrained, i.e., not every amino acid can be substituted for any other amino acid. To find out whether such constraints could explain the extent of cycles in the genotype network, I first quantified these constraints. Among the 329 amino acids in the protein coding sequence, 156 vary in the present data set. On average, each of these 156 sites is involved in 3.5 change events (edges), but with a distribution that is broad and ranges from one edge to 16 edges (Figure S18a). The total number of different amino acids that occurs at each variable position ranges from two to four (Figure S18b), with a mean of 2.67. Positions that are involved in more change events also harbor significantly more amino acids (Spearman's $R=0.74$, $n=156$; $P<10^{-17}$).

Next I created random genotype networks that reflect these constraints in a conservative way. That is, they are based on sequences with 156 variable positions, but I assumed that each position can only admit two different amino acids. The distribution of the number of sequences that form cycles in 1000 such random networks is shown in Figure S18c (note the logarithmic vertical scale). 905 of these networks contain no cycle at all, and the maximum number of sequences in cycles is twelve for any of the networks. Not a single random network contains 79 sequences in cycles, as does the largest component of the actual network. These observations suggest that the abundance of cycles in the genotype network cannot be explained from global constraints on sequence evolution.

Squares and triangles reflect two different kinds of constrained evolution. All triangles involve change at only a single amino acid site (Table S1), as in the example of the triangle between strains Miyazaki/39/2005, Nagoya/26/2006, and Osaka/4/2006 (Figure 2c) which differ only in position 222. If

the Miyazaki strain is the ancestor of the other two strains, then the arginine (R) at its position 222 gave rise to a lysine (K) in the Nagoya strain and an isoleucine (I) in the Osaka strain. Analogous scenarios hold if either the Nagoya or the Osaka strain are ancestral. This triangular pattern of change reflects strongly constrained sequence evolution: Out of all possible positions that could have changed in the common ancestor, only one did change.

Figure 2d illustrates that squares also reflect a pattern of constrained evolution, but of a different kind that leads to temporary sequence convergence. If Mae Hong Son/33/2003 is the most recent common ancestor of its two neighboring strains, then it underwent a valine to isoleucine substitution at position 226 (I226V) that created strain Singapore/95/2003, as well as a tyrosine to phenylalanine substitution at position 159 (Y159F) that produced strain Ukraine/UA-2003918069/2003. Thus, unlike in a triangle, two different positions changed in the ancestor. Then either the Singapore strain underwent the same Y159Y change involved in creating the Lipetsk strain, or the Ukraine strain underwent a I226V change (or both kinds of changes occurred). Regardless of this order, and regardless of which strain is the common ancestor of the others, the characteristic pattern is that its descendants temporarily become more similar to one another, such as the Lipetsk strain which differs in only one amino acid from the Ukraine strain, whereas its Singaporean ancestor differs in two amino acids from the Ukraine strain. This temporary sequence convergence characterized by the same two substitutions on opposing edges of a square (Figure 2d) exists for all but one square (Table S2). This only exception is a square where sequence change occurred at only one site (Figure S7), but this square is really just a composite of two triangles. I note that a square cannot involve substitutions at more than two positions, because that would make cycle-closure after four edges impossible. Figure 2e illustrates how the group of four strains from Figure 2d would appear in a maximum-likelihood phylogenetic tree of the HA sequences considered here (Figure S13). The four-strain subtree correctly captures the single amino acid change that separates the pairs of isolates Lipetsk-Ukraine, Lipetsk-Singapore, and Singapore-Mae Hong Son, but it incorrectly suggests that Lipetsk-Mae Hong Son are two amino acid changes apart, whereas they actually differ in only one amino acid. Other instances of squares would show the same misleading phylogenetic topology.

Cycles are enriched with changes in epitopes. HA sequences are subject to positive selection of beneficial mutations (Bhatt et al. 2011; Bush et al. 1999a; Bush et al. 1999b; Suzuki 2006; Suzuki ; Wolf et al. 2006) that frequently occur in antibody-binding epitopes whose mutational change helps a virus evade the host's immune response (Munoz & Deem 2005; Wiley et al. 1981). Such changes are also involved in the HA genotype network, because amino acid change in this network is almost twice as likely to occur in an epitope than outside an epitope (Figure S8; $P<4x10^{-3}$). More importantly, epitope-affecting changes are especially prevalent in cycles. That is, significantly more amino acid changes in cycles affect epitopes than outside cycles. This difference also holds for triangles and squares when they are considered separately (Supplementary Results).

Tolerable or beneficial mutations are especially rare in cycles. A frequently used direct indicator of positive selection is the ratio d_N/d_S of nonsynonymous changes d_N per non-synonymous site and synonymous changes d_S per synonymous site(Bush et al. 1999b; Kryazhimskiy et al. 2008; Kryazhimskiy et al. 2011; Kryazhimskiy & Plotkin 2008; Suzuki 2006; Suzuki 2008; Suzuki & Gojobori 1999; Wolf et al. 2006). Using it to identify whether individual amino acid changes have been beneficial is not

straightforward. First, only in highly divergent sequences does a ratio d_N/d_S >1 indicate positive selection (Bush et al. 1999b; Kryazhimskiy & Plotkin 2008). In sequences with low divergences like those considered here, positive selection can be at work even if d_N/d_S is significantly smaller than one. Second, neighboring sequences in the genotype network differ in only a single amino acid and often show no synonymous change (Figure S9), such that this ratio cannot even be determined for many individual amino acid changes. These considerations show that any analysis of d_N/d_S needs to focus on *groups* of edges in this genotype network. If one considers all amino acid changes in the network, one finds more than 100 edges (117 of 539, 21.7 percent) in the genotype network where an amino acid change occurred without any synonymous change, suggesting that positive selection occurred at least in some of the sequences considered here.

In closely related sequences, the relative incidence of synonymous changes to amino acid changes can be useful to indicate the average time between occurrences of tolerable or beneficial amino acid changes. Many synonymous changes per amino acid change indicates a long waiting time. Under neutral evolution, the expected number of synonymous changes per amino acid change in the data set considered here is approximately 0.36 (±0.02 s.e.m; see Methods). The mean number of synonymous changes per amino acid change in the whole network is much higher at 2.12 (±0.08 s.e.m, n=539). What is more, this number is even higher for edges in cycles (2.51 ±0.16 s.e.m, n=180), and it is lower for all edges outside cycles (1.93 ±0.09 s.e.m, n=359), a difference that is significant (p=0.0089, Mann-Whitney U-test, Figure S10a). Even in the edge with the lowest synonymous distance within any one cycle, this distance is significantly greater than that of comparable edges outside cycles (Supplementary Results, Figure S10). In sum, tolerable or beneficial mutations are rarer in cycles than in the rest of the network. An additional pertinent observation is that cycles are enriched with codons experiencing amino acid change through double or triple nucleotide change, and the HA sequences in which these codons reside also show greater synonymous divergence (Supplementary results, Figure S11, Figure S12).

Highly central strains and bridge strains reflect the trunk-like genealogy of HA sequences. Phylogenetic trees of evolving HA sequences have a trunk-like structure: Relatively few sequences in the trunk propagate the lineage further, whereas sequences in side branches are evolutionary dead-ends (Bedford et al. 2012; Koelle et al. 2006; Wolf et al. 2006; Zinder et al. 2013). This topology is responsible for a disassortative network organization (Figure S2). One can quantify the "trunkness" or centrality of a sequence through the graph-theoretical concept of a node's betweenness centrality *B* – the number of shortest paths connecting pairs of nodes that pass through that node. Figure 3a shows the largest component of the HA genotype network with the most central strains (Table S3) are highlighted. The most central strain is Taiwan/TW-1554/2004, with B=20,020 shortest paths passing through it, followed by Okayama/15/2005 (B=17,528 paths) and Osaka/18/2006 (B=15,213 paths). These strains form part of the trunk of the HA phylogeny, giving rise to many descendants in the HA phylogenetic tree (Figure S13).

The more surviving descendants a virus leaves – the higher its number of neighbors in a genotype network – the greater should be the likelihood that it is a trunk strain, because chances are greater that one of its descendants propagates the lineage further. The quantitative analysis of Figure 3b supports this assertion by showing that strains with many neighbors are significantly more central (Spearman's r=0.91,p $<10^{-17}$, n=246). But more remarkable than this rule are its exceptions, because several central

strains have few immediate neighbors. These include the strains Osaka/18/2006, ranked third in terms of centrality, but having only 6 neighbors, as well as Lipetsk/15/2004 (rank 5, 8 neighbors), Hong-Kong/2982/2004 (rank 9, 2 neighbors), Perth/20/2005 (rank 10, 3 neighbors). Because such unusual strains visibly link major clusters of sequences, I refer to them as *bridge strains* (see also Table S3). That all these bridge strains are mere artefacts of undersampling in time or space is possible, but made less likely by the observation that (i) sampling in time for the sequences shown in Figure 3a is quite even, ranging from 113 isolates in 2002 to 173 isolates in 2006, and (ii) three of the four bridge strains above come from the top four (out of 53) sampled countries Japan, China, and Australia. That such strains bear some biological significance is also made more by the observation that highly central and bridge strains ($B>2500$ and degree≤ 10) are significantly enriched in cycles and, more specifically, squares (Figures 3c and 3d; $P<10^{-17}$; Mann-Whitney U-test). Such enrichment does not exist for triangles ($P>0.29$). Central and bridge strains thus show an elevated incidence of convergent evolution (See supplementary results for their possible biological significance).

Discussion

The HA genotype network differs in a major respect from phylogenetic trees – those acyclic graphs used to describe groups of sequences related by descent – because cycles permeate this network. For example, the diameter of the subgraph formed by those strains that are part of a cycle (Figure 2a) equals nine edges, only one fewer than the whole network's diameter of 10 edges, which implies that one can traverse almost the entire network along cycles. Most of these cycles are squares, which reflects an extent of sequence homoplasy that can neither be explained by chance alone, nor by global constraints on sequence evolution (Figure S18). All observed cycles are very short and involve change at only one or two sites, which suggests that sequences in a cycle can experience very few tolerable or beneficial amino acid changes. This is further underscored by the observation that some amino acid sites are involved in multiple homoplastic cycles, such as site 50, which is involved in 7 different squares (Table S2). The limited overlap between sites that undergo homoplastic change here and in previous studies further highlight this sequence context specificity (Kryazhimskiy et al. 2008; Kryazhimskiy et al. 2011; Wolf et al. 2006). For example, among the 25 sites reported by (Kryazhimskiy et al. 2008) to have undergone directional evolution, a form of homoplasy, only 13 are implicated in such change here, and among 11 sites reported by (Wolf et al. 2006), only 2 are implicated in this data set. (All of these common sites are part of an epitope.) More generally, the genotype network approach reveals the extent of homoplasy to be more extreme than hitherto realized. Previous analyses indicated homoplastic changes between leaf sequences on different branches of a HA phylogenetic tree, usually separated by multiple further additional amino acid changes (Kryazhimskiy et al. 2008; Kryazhimskiy et al. 2011; Wolf et al. 2006), but convergent changes in the present data set occur in the same square, only one amino acid change apart. I note that the amino acid changes that occur in squares are suggestive of epistatic interactions (Bershtein et al. 2006; Bonhoeffer et al. 2004; Cordell 2002; Kryazhimskiy et al. 2011; Kulathinal et al. 2004; Wilke et al. 2003), which demonstrably exist in influenza HA evolution (Kryazhimskiy et al. 2011) and can be a source of homoplasy.

Is the observed homoplasy only caused by extremely few tolerable amino acid changes, or do positive selection and beneficial mutations contribute to it? Positive selection is pervasive in HA evolution (Bhatt

et al. 2011; Bush et al. 1999a; Bush et al. 1999b; Huang & Yang 2011; Kryazhimskiy et al. 2008; Suzuki 2006; Suzuki 2008; Wolf et al. 2006). Unfortunately, one prominent criterion to detect it, a ratio $d_N/d_S>1$ is only valid for sequences much more distantly related than those considered here (Kryazhimskiy & Plotkin 2008), whose pairwise nucleotide divergence of $3.22x10^{-3}$ is similar to that within a single influenza outbreak ($3.4x10^{-3}$, Lavenu et al. 2006). For such lowly diverged sequences that coexist in the same population, the ratio d_N/d_S can be significantly lower than one, yet positive selection may be rampant (Kryazhimskiy & Plotkin 2008). However, at least some of the homoplastic changes observed here are likely to be beneficial, because, first, 32 of 117 edges (27.3 percent) that involve no synonymous change at all occur in cycles. Second and more importantly, amino acid changes in cycles are significantly more likely to affect epitopes than changes outside cycle (Supplementary Results). Given past observations on the association of beneficial changes with epitopes (Bush et al. 1999b; Huang & Yang 2011; Kryazhimskiy et al. 2008; Wolf et al. 2006), this observation is not consistent with the notion that homoplasy occurs only because of increased selective constraints.

In principle, the rate of antigenic evolution in influenza could be limited by the rarity of mutations that cause antigenic change (Koelle et al. 2006; Russell et al. 2008; Wolf et al. 2006), or by immune-mediated selection in the host. In the latter case, only a limited number of antigenic types exists, which circulate in a population. Temporally changing patterns of host immunity can then cause different types to rise to prominence over time (Recker et al. 2007; Wikramaratna et al. 2013). The present data can speak to the question whether mutation limitation contributes to HA evolution, because the ratio d_N/d_S can help estimate the waiting times between successive amino acid changes: If synonymous changes occur according to a molecular clock, as is the case for influenza (Jenkins et al. 2002), then a higher number of synonymous changes per amino acid change imply a longer waiting time. It is of particular interest to study changes in cycles, because such changes are significantly associated with epitopes (Supplementary Results). Based on the significantly higher average number of synonymous changes inside than outside cycles (Figure S10, 2.51 versus 1.93, or $2.54x10^{-3}$ and $1.96x10^{-3}$ per nucleotide site), and given an estimated 3.4 (±1.1x10-3, Jenkins et al. 2002) silent changes per site per year, the expected waiting time between two amino acid changes is 273 days inside cycles and 210 days outside cycles. Moreover, the HA sequences with non-synonymous codons that carry double nucleotide changes have a higher number of 3.17 (±0.38 s.e.m.) synonymous changes than HA sequences where these codons experienced only a single nucleotide change (2.05±0.08 s.e.m.), a difference that is significant ($p=0.002$; Mann-Whitney U-test) and that amounts to an increase in the expected waiting time of 54 percent (from 223 to 344 days). These observations suggest that mutation-limitation plays some role in HA evolution. The episodic occurrence of beneficial mutations (Koelle et al. 2006; Nelson & Holmes 2007; Wolf et al. 2006) and results from recent epidemiological modeling (Koelle et al. 2006; Zinder et al. 2013) also support this notion.

Genotype networks and phylogenetic trees are graphs with complementary strengths for the analysis of evolutionary data. First, phylogenetic analysis uses sophisticated algorithms (Felsenstein 2004) to infer a tree's structure from data on its leaf nodes, whereas the structure of a genotype network immediately follows from the raw data. Second, in a tree the genotype of interior (ancestral) nodes need to be inferred probabilistically (Felsenstein 2004; Guindon & Gascuel 2003), whereas in a genotype network

these nodes are plainly visible. Third, a phylogenetic tree has inherent ancestor-descendant directionality, which would need to be inferred in a genotype network. Fourth and conversely, although phylogenetic analysis can detect homoplasy (Delport et al. 2008; Kryazhimskiy et al. 2008; Wolf et al. 2006), homoplasy is a confounding factor in tree reconstruction and not readily visible from a tree itself, whereas cycles make homoplasy plainly visible in a genotype network. Fifth and finally, trees are well-suited to study evolutionary relationships of sequences with arbitrarily high divergence, whereas genotype networks would be highly fragmented and thus of limited use for such sequences. But once abundant data on closely related sequences are available, genotype networks become highly useful tools to understand evolutionary processes at fine-grained temporal resolution.

Materials and Methods

I obtained DNA sequences and the amino acid sequences they encode for 1565 influenza A haemagglutinin (HA) genes from the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov), using database accession numbers published in Table S1 of (Russell et al. 2008). These sequences come from influenza A (H3N2) isolates obtained world-wide between 2002 and 2007. Their antigenic properties, identified through hemagglutinin inhibition assays, were reported and analyzed in (Russell et al. 2008). I aligned DNA and amino acid sequences with the program MUSCLE (Edgar 2004) using default parameters. The resulting alignments contained no gaps. To build a genotype network of amino acid sequences from this alignment, *I* defined a graph (*V*,*E*) whose set of nodes *V* comprises all genotypes (amino acid sequences), and where two nodes *v* and *w* are connected by an edge (*v*,*w*) in the set of edges *E* if they differ in exactly one amino acid. A cycle is a sequence of edges in a graph that returns to the sequence's starting node, but does not visit any node or edge more than once. I computed the number of nodes that are part of a cycle, and performed all other graph computations and visualizations with the aid of the perl package GRAPH (version 0.94; www.cpan.org), as well as with gephi (version 0.8.2 Bastian et al. 2009). I associated each hemagglutinin sequence in the network with information about its year and country of isolation, which I extracted from the annotated sequence files in Table S1 of (Russell et al. 2008), and analyzed categorical data with the aid of the R package 'vcd'. Methods are described in greater detail in the Supplementary Methods.

Acknowledgments

I would like to thank members of the A*star Bioinformatics in Singapore, and especially Dr. Sebastian Maurer-Stroh, for valuable discussions during a sabbatical visit in Singapore. This work has been partially supported through Swiss National Science Foundation grant 315230-129708, as well as by the URPP Evolutionary Biology at UZH.

Data accessibility

This work uses only publicly accessible data (see Methods).

Figure Legends

Figure 1: Protein genotype network without redundant sequences. a) Component size distribution of the HA protein genotype network, where groups of identical amino acid sequences are represented by only one member. **b)** The largest connected component, where circles correspond to sequences, and edges connect sequences that differ in a single amino acid. The component's layout is computed by a force-directed algorithm (Hu 2005). Larger circles and darker hues of blue correspond to sequences with higher degree (more neighbors). **c)** Degree distribution of the entire genotype network. The left-most bar in the histogram includes isolated nodes with no neighbors as well as nodes with one neighbor. Note the many low-degree nodes.

Figure 2: The protein genotype network contains many cycles. a) The 2-core of the protein genotype's largest component shown in Figure 1. Apparent differences between the local topologies of the two graphs are a consequence of the embedding algorithm (Hu 2005). The 2-core of the entire genotype network contains all 122 nodes that are part of a cycle. Larger circles and darker hues of blue correspond to sequences with higher degree (more neighbors). **b)** Venn diagram showing the number of nodes contained in triangles (48=28+20 nodes, left ellipse), in squares (94=20+74 nodes), and in both (20 nodes).These numbers refer to the entire protein genotype network, not just the largest component. **c)** An example of a triangle in the protein genotype network**.** Circles correspond to HA sequences from strains whose names are shown above each circle. Edges are labeled with the amino acid difference between two neighboring strains, e.g., K222I indicates that strain Nagoya/26/2006 contains lysine (K) at position 222, whereas strain Osaka/4/2006 contains isoleucine (I) at that position of the HA amino acid sequence. **d)** A square in the genotype network. Note that amino acids at two positions (159 and 226) change in this square. The amino acid differences are read from the left node to the right node in each sequence. **e)** The topology of a subtree of the maximum-likelihood HA phylogeny from Figure S13, containing the four different influenza isolates from Figure 2d. Each tree branch connecting two nodes corresponds to exactly a single amino acid change (dashed arrow). The dashed line at the root would connect this clade to the much larger HA sequence tree.

Figure 3: Central sequences in the genotype network include some bridge sequences with few neighbors. a) The largest connected component of the genotype network, where circles correspond to sequences, and edges connect sequences that differ in a single amino acid. The topology of this network is the same as that of Figure 1b, apparent differences being a consequence of the embedding algorithm (Hu 2005). Circles in purple correspond to those 11 sequences in the largest component of the protein genotype network whose betweenness centrality exceeds 2500, i.e., those sequences through which the most shortest paths between other sequences must pass. The component's layout is computed by a force-directed algorithm (Hu 2005). **b)** Scatterplot of betweenness centrality B for strains where B>0 (horizontal axis) versus their number of neighbors (vertical axis). Note the logarithmic scale on each axis. A few strains have high betweenness centrality but relatively few neighbors. The names of three such bridge strains are written immediately underneath the three circles representing their centrality and degree. **c)** Mean (±s.e.m.) of betweenness centrality for nodes that are part of a cycle (left) and not part of a cycle (right). **d)** Mean (±s.e.m.) of betweenness centrality are part of a square (left) and not part of a

cycle (right). The differences are highly significant in both cases ($P<10^{-17}$; Mann-Whitney U-test), showing that nodes in squares tend to be central.

References

Bascompte, J. & Jordano, P. 2007 Plant-animal mutualistic networks: The architecture of biodiversity. In *Annual Review of Ecology Evolution and Systematics*, vol. 38, pp. 567-593.

Bascompte, J., Jordano, P. & Olesen, J. M. 2006 Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science* **312**, 431-433.

Bastian, M., Heymann, S. & Jacomy, M. 2009 Gephi: an open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media.*

Bedford, T., Rambaut, A. & Pascual, M. 2012 Canalization of the evolutionary trajectory of the human influenza virus. *Bmc Biology* **10**.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. 2006 Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929-932.

Bhatt, S., Holmes, E. C. & Pybus, O. G. 2011 The genomic rate of molecular adaptation of the human influenza A virus. *Molecular Biology and Evolution* **28**, 2443-2451.

Bollobás, B. 1985 *Random graphs*. London: Academic Press.

Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M. & Petropoulos, C. J. 2004 Evidence for positive epistasis in HIV-1. *Science* **306**, 1547-1550.

Bull, P. C., Buckee, C. O., Kyes, S., Kortok, M. M., Thathy, V., Guyah, B., et al. 2008 Plasmodium falciparum antigenic variation. Mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Molecular Microbiology* **68**, 1519-1534.

Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. 1999a Predicting the evolution of human influenza A. *Science* **286**, 1921-1925.

Bush, R. M., Fitch, W. M., Bender, C. A. & Cox, N. J. 1999b Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution* **16**, 1457-1465.

Carrat, F. & Flahault, A. 2007 Influenza vaccine: The challenge of antigenic drift. *Vaccine* **25**, 6852-6862.

Ciliberti, S., Martin, O. C. & Wagner, A. 2007 Circuit topology and the evolution of robustness in complex regulatory gene networks. *PLoS Computational Biology* **3(2): e15**.

Cohen, J. E. & Briand, F. 1984 Trophic Links of Community Food Webs. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **81**, 4105-4109.

425 Colizza, V., Barrat, A., Barthelemy, M. & Vespignani, A. 2006 The role of the airline transportation
426 network in the prediction and predictability of global epidemics. *Proceedings of the National*
427 *Academy of Sciences of the United States of America* **103**, 2015-2020.
428 Cordell, H. J. 2002 Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in
429 humans. *Human Molecular Genetics* **11**, 2463-2468.
430 Delport, W., Scheffler, K. & Seoighe, C. 2008 Frequent toggling between alternative amino acids Is driven
431 by selection in HIV-1. *PLoS Pathogens* **4(12)**, e1000242.
432 Edgar, R. C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic*
433 *Acids Research* **32**, 1792-1797.
434 Felsenstein, J. 2004 *Inferring phylogenies*. Sunderland, Massachusetts: Sinauer Associates.
435 Guindon, S. & Gascuel, O. 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by
436 maximum likelihood. *Systematic Biology* **52**, 696-704.
437 Harary, F. 1969 *Graph theory*. Reading, Massachusetts: Addison-Wesley.
438 Hayden, E. J., Ferrada, E. & Wagner, A. 2011 Cryptic genetic variation promotes rapid evolutionary
439 adaptation in an RNA enzyme. *Nature* **474**, 92-U120.
440 Hayward, A. C., Fragaszy, E. B., Bermingham, A., Wang, L., Copas, A., Edmunds, W. J., et al. 2014
441 Comparative community burden and severity of seasonal and pandemic influenza: results of the
442 Flu Watch cohort study. In *The Lancet Respiratory Medicine*. doi:10.1016/S2213-2600(14)70034-
443 7
444 Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. 2011 Experimental illumination of a fitness landscape.
445 *Proceedings of the National Academy of Sciences of the United States of America* **108**, 7896-
446 7901.
447 Hu, Y. 2005 Efficient and high quality force-directed graph drawing. *Mathematica Journal* **10**, 37-71.
448 Huang, J. W. & Yang, J. M. 2011 Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses.
449 *Bmc Bioinformatics* **12**.
450 Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. 2002 Rates of molecular evolution in RNA
451 viruses: A quantitative phylogenetic analysis. *Journal of Molecular Evolution* **54**, 156-165.
452 Koelle, K., Cobey, S., Grenfell, B. & Pascual, M. 2006 Epochal evolution shapes the phylodynamics of
453 interpandemic influenza A (H3N2) in humans. *Science* **314**, 1898-1903.
454 Kryazhimskiy, S., Bazykin, G. A., Plotkin, J. & Dushoff, J. 2008 Directionality in the evolution of influenza A
455 haemagglutinin. *Proceedings of the Royal Society B-Biological Sciences* **275**, 2455-2464.

456 Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. 2011 Prevalence of epistasis in the evolution of
457 influenza A surface proteins. *Plos Genetics* **7**.
458 Kryazhimskiy, S. & Plotkin, J. B. 2008 The population genetics of dN/dS. *Plos Genetics* **4**.
459 Kulathinal, R. J., Bettencourt, B. R. & Hartl, D. L. 2004 Compensated deleterious mutations in insect
460 genomes. *Science* **306**, 1553-1554.
461 Lavenu, A., Leruez-Ville, M., Chaix, M. L., Boelle, P. Y., Rogez, S., Freymuth, F., et al. 2006 Detailed
462 analysis of the genetic evolution of influenza virus during the course of an epidemic.
463 *Epidemiology and Infection* **134**, 514-520.
464 Lipman, D. & Wilbur, W. 1991 Modeling neutral and selective evolution of protein folding. *Proceedings of*
465 *the Royal Society of London Series B* **245**, 7-11.
466 Maslov, S. & Sneppen, K. 2002 Specificity and stability in topology of protein networks. *Science* **296**, 910-
467 913.
468 Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. 2002 Network motifs: Simple
469 building blocks of complex networks. *SCIENCE* **298**, 824-827.
470 Munoz, E. T. & Deem, M. W. 2005 Epitope analysis for influenza vaccine design. *Vaccine* **23**, 1144-1148.
471 Nelson, M. I. & Holmes, E. C. 2007 The evolution of epidemic influenza. *Nature Reviews Genetics* **8**, 196-
472 205.
473 Newman, M. E. J. 2002 Assortative mixing in networks. *Physical Review Letters* **89**.
474 Newman, M. E. J. 2006 Modularity and community structure in networks. *Proceedings of the National*
475 *Academy of Sciences of the United States of America* **103**, 8577-8696.
476 Onnela, J. P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., et al. 2007 Structure and tie
477 strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*
478 *of the United States of America* **104**, 7332-7336.
479 Pybus, O. G. & Rambaut, A. 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nature*
480 *Reviews Genetics* **10**, 540-550.
481 Recker, M., Pybus, O. G., Nee, S. & Gupta, S. 2007 The generation of influenza outbreaks by a network of
482 host immune responses against a limited set of antigenic types. *Proceedings of the National*
483 *Academy of Sciences of the United States of America* **104**, 7711-7716.
484 Rodrigues, J. F. M. & Wagner, A. 2009 Evolutionary plasticity and innovations in complex metabolic
485 reaction networks. *PLoS Computational Biology* **5**.
486 Romero, P. A. & Arnold, F. H. 2009 Exploring protein fitness landscapes by directed evolution. *Nature*
487 *Reviews Molecular Cell Biology* **10**, 866-876.

488 Rozenfeld, A. F., Arnaud-Haond, S., Hernandez-Garcia, E., Eguiluz, V. M., Serrao, E. A. & Duarte, C. M.
489 2008 Network analysis identifies weak and strong links in a metapopulation system. *Proceedings*
490 *of the National Academy of Sciences of the United States of America* **105**, 18824-18829.
491 Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., et al. 2008 The global circulation
492 of seasonal influenza A (H3N2) viruses. *Science* **320**, 340-346.
493 Schuster, P., Fontana, W., Stadler, P. & Hofacker, I. 1994 From sequences to shapes and back - a case-
494 study in RNA secondary structures. *Proceedings of the Royal Society of London Series B* **255**, 279-
495 284.
496 Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A., et al.
497 2004 Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371-376.
498 Suzuki, Y. 2006 Natural selection on the influenza virus genome. *Molecular Biology and Evolution* **23**,
499 1902-1911.
500 Suzuki, Y. 2008 Positive selection operates continuously on hemagglutinin during evolution of H3N2
501 human influenza A virus. *Gene* **427**, 111-116.
502 Suzuki, Y. & Gojobori, T. 1999 A method for detecting positive selection at single amino acid sites.
503 *Molecular Biology and Evolution* **16**, 1315-1328.
504 Wagner, A. 2011 *The origins of evolutionary innovations. A theory of transformative change in living*
505 *systems.* Oxford, UK: Oxford University Press.
506 Wagner, R., Matrosovich, M. & Klenk, H. D. 2002 Functional balance between haemagglutinin and
507 neuraminidase in influenza virus infections. *Reviews in Medical Virology* **12**, 159-166.
508 World Health Organization. 2009 Influenza (seasonal). In *WHO fact sheet No. 211*.
509 Wikramaratna, P. S., Sandeman, M., Recker, M. & Gupta, S. 2013 The antigenic evolution of influenza:
510 drift or thrift? *Philosophical Transactions of the Royal Society B-Biological Sciences* **368**.
511 Wiley, D. C., Wilson, I. A. & Skehel, J. J. 1981 Structural identification of the antibody-binding sites of
512 Hong-Kong influenza hemagglutinin and their involvement in antigenic variation. . *Nature* **289**,
513 373-378.
514 Wilke, C. O., Lenski, R. E. & Adami, C. 2003 Compensatory mutations cause excess of antagonistic
515 epistasis in RNA secondary structure folding. *BMC Evolutionary Biology* **3**, 3.
516 Wilson, I. A., Skehel, J. J. & Wiley, D. C. 1981 Structure of the hemagglutinin membrane glycoprotein of
517 influenza virus at 3 A resolution. *Nature* **289**, 366-373.
518 Wolf, Y. I., Viboud, C., Holmes, E. C., Koonin, E. V. & Lipman, D. J. 2006 Long intervals of stasis punctuated
519 by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct* **1**.

520 Zinder, D., Bedford, T., Gupta, S. & Pascual, M. 2013 The roles of competition and mutation in shaping
521 antigenic and genetic diversity in influenza. *Plos Pathogens* **9**.

522

523